確率一般化LR構文解析の先読み方式変更による拡張 (計算機科学の基礎理論 : 21世紀の計算パラダイムを目指して)

AUTHOR(S):

椎名, 広光; 増山, 繁

CITATION:

椎名, 広光 ...[et al]. 確率一般化LR構文解析の先読み方式変更による拡張 (計算機科学の基礎理論 : 21世紀の計算パラダイムを目指して). 数理解析研究所講究録 2000, 1148: 170-174

ISSUE DATE:

2000-04

URL:

http://hdl.handle.net/2433/63995

数理解析研究所講究録
1148巻 2000年 170-174

確率一般化LR構文解析の先読み方式変更による拡張

椎名 広光 (Hiromitsu Shiina): 岡山理科大学 総合情報学部
増山 繁 (Shigeru Masuyama): 豊橋技術科学大学 知識情報工学系

1 まえがき

確率構文解析法は，音声認識における音素の予測モデルなどの統計的言語モデルとして用いられてきている．これは文脈自由文法の生成規則に出現確率を付加したモデルで, inside-outside アルゴリズムなどによって生成規則を推定し，尤もらしい構文解析木を求めるモデルが知られている．また，プッシュダウンオートマトンを予め作成しておいて，それを用いて構文解析を進めるLR構文解析法を，文脈自由言語に対する構文解析法に拡張した一般化LR構文解析法 [1](以降，GLR構文解析) が知られている．

これに対しBriscoeとCarrollは，GLR構文解析法を応用して状態と先読み文字列の組合せに対して出現確率を付加する確率GLR構文解析法 [2](以降，B&C法と呼ぶ) を提案している．しかしながら，出現確率の正規化に問題があり，構文解析木の正しい生起確率を求めることができない場合があった．それに対し，Inuiら [2] によって提案されている確率GLR構文解析 (以降，Inui法と呼ぶ) では，出現確率の情報を持つ状態と先読み文字列の組合せの方法を工夫することによって，B&C法で起こる出現確率の正規化の問題を解決し，構文解析木の生起確率の近似値の精度を改善している．

ところが，学習させる構文解析木が複雑な場合は，B&C法やInui法においては，現在の状態と先読み文字列及び次に予測される状態の組合せだけでは，構文解析木の正しい生起確率を計算することができない．

そこで，本稿では，入力データから一般化LR構文解析によって構文解析木を抽出することができることを前提とし，この時構文解析木の正しい生起確率を求める問題において，学習させる構文解析木を用いて出現確率を，次の入力文字，これから遷移する先読み状態 (以降，先読み状態)，構文解析中にスタックに積まれている状態ごとに求めるように拡張する．その結果，構文解析途中のスタックに積まれている状態の個数と先読み状態の個数が十分大きければ，拡張した確率一般化LR構文解析により構文解析木の正しい生起確率を求めることができることを示す．また，スタックに積まれている状態数と先読みの状数数の削減方法についても述べる．

2 準備

与えられた文脈自由文法 $G = (P, N, T, S, \$)$, P: 生成規則の集合, N: 非終端記号の集合, T: 終端記号, S: 開始記号，$\$$: 入力の終りを示す特殊記号から作られるLR状態遷移図 $\mathbf{LRA} = (\mathbf{S}, \delta)$ は次のように定義される．

(1) $\mathbf{S}$: LR状態遷移図の状態集合．各状態における動作はプッシュダウンオートマトンの動作を表すため，状態集合は **shift** 状態集合と **reduce** 状態集合と受理を表す状態 s_{acc} に分けられ，$\mathbf{S} = \mathbf{S_s} \cup \mathbf{S_r} \cup \{s_{acc}\}$ となる．また，開始状態を s_0 と表し，$s_0 \in \mathbf{S_s}$ である．

(2) δ: LR状態遷移図の各状態における動作を表す．状態は **shift** 状態と **reduce** 状態に分けられる．**shift** 状態では，**shift** 状態と文法記号の組合せに対して，次に遷移する先の状態を関数 $\delta_\mathbf{s}$ で表す．

$$\delta_\mathbf{s}(s_i, l_i) = \{s_j \mid s_i, s_j \in \mathbf{S_s}, l_i \in T \cup N\}.$$

一方，**reduce** 状態では，**reduce** 状態で還元する文法規則の種類を関数 $\delta_\mathbf{r}$ で表す．

$$\delta_\mathbf{r}(s_i) = \{A \to \alpha \mid s_i \in \mathbf{S_r}, \text{“}A \to \alpha\text{”} \in P\}.$$

$\delta = \delta_\mathbf{s} \cup \delta_\mathbf{r}$ である．

また，ここで述べたLR状態遷移図は表の形で表した場合，表1のようになる．LR表の **re** は δ_r の遷移に対応し，それ以外は δ_s の遷移に対応する．なお，LR構文解析の動作などの詳細は，文献 [4] などを参照されたい．

3 一般化LR構文解析法

本節では，一般化LR構文解析法のうちInui法について述べる．

3.1 Inui 法

Inui 法では，入力文字列 $l_1 l_2 \ldots l_n$ による開始状態 s_0 から受理状態 s_{acc} までの遷移の過程における状態を積んでいるスタックの内容 (以下，簡単のため，混乱を生じない限り，単に「スタック」と表す) を一つの「状態」と考えている．すなわち，$\sigma_i \in \sigma$ を時点 i におけるスタックの内容とし，入力文字 l_i，動作 a_i とすると，「状態」遷移過程は次のように示される．

$$\sigma_0 \overset{l_1,a_1}{\Longrightarrow} \sigma_1 \overset{l_2,a_2}{\Longrightarrow} \cdots \overset{l_{n-1},a_{n-1}}{\Longrightarrow} \sigma_{n-1} \overset{l_n,a_n}{\Longrightarrow} \sigma_n$$

これに対して，構文解析木 T の生起確率 $P(T)$ も同様にスタックを用いて定義している．また，時点 i における入力文字 l_i，動作 a_i，スタック σ_i は時点 $i-1$ におけるスタック σ_{i-1} に依存すると仮定すると，次のように定義している．

$$\begin{aligned} P(T) &= P(\sigma_0, l_1, a_1, \sigma_1, \ldots, l_n, a_n, \sigma_n) \\ &= \prod_{i=1}^{n} P(l_i, a_i, \sigma_i \mid \sigma_0, l_1, a_1, \sigma_1, \ldots, \\ &\qquad l_{i-1}, a_{i-1}, \sigma_{i-1}) \\ &\approx \prod_{i=1}^{n} P(l_i, a_i, \sigma_i \mid \sigma_{i-1}) \end{aligned}$$

$P(l_i, a_i, \sigma_i \mid \sigma_{i-1})$ では，全てのスタックを保持しておかなければならないので，スタック σ_i の先頭の状態 s_i で近似している．

また，時点 i における 1 つ前のスタック先頭の状態 s_i が **shift** 状態であるかまたは **reduce** 状態であるかによって条件付確率を変更している．この変更が B&C 法の方法で起こる出現確率の正規化の問題を Inui 法が解決している点である．

$$\begin{aligned} &P(l_i, a_i, \sigma_i \mid \sigma_{i-1}) \\ &\quad \approx \begin{cases} P(l_i, a_i \mid s_i), & \text{if } a_i = \textbf{shift} \\ P(a_i \mid s_{i-1}, l_i), & \text{if } a_i = \textbf{reduce} \end{cases} \end{aligned}$$

4 拡張 PGLR

本論文で提案する方法では，構文解析時のスタックの状態と先読みとして状態遷移を用いる条件付確率で，構文解析木 T の生起確率 $P(T)$ を近似する．

本論文では，構文解析時のスタックの状態と先読みとして状態遷移を用いる条件付確率で，構文解析木 T の生起確率 $P(T)$ を近似する方法を提案する．

表 1: 文法 G_1 の LR 表

state	action				goto
	u	v	x	\$	S
S_0	**sh1**	**sh2**	**sh3**		S_4
S_1	**sh1**	**sh2**	**sh3**		S_5
S_2	**sh1**	**sh2**	**sh3**		S_6
S_3	**re3**	**re3**	**re3**		
S_4	**re3**	**re3**	**re3**	**acc**	S_7
S_5	**sh1/re1**	**sh2/re1**	**sh3/re1**	**re1**	S_7
S_6	**sh1/re2**	**sh2/re1**	**sh3/re1**	**re2**	S_7
S_7	**sh1/re4**	**sh2/re4**	**sh3/re4**	**re4**	S_7

$$\begin{aligned} P(T) &= P(\sigma_0, l_1, a_1, \sigma_1, \ldots, l_n, a_n, \sigma_n) \\ &= \prod_{i=1}^{n-1} P(\sigma_i, l_i, a_i, \sigma_{i+1}, \ldots, l_n, a_n, \sigma_n \mid \\ &\qquad \sigma_0, l_1, a_1, \sigma_1, \ldots, l_{i-1}, a_{i-1}, \sigma_{i-1}) \\ &= \prod_{i=1}^{n-1} P(l_i, a_i, \sigma_i \sigma_{i+1} \ldots \sigma_n \mid \sigma_{i-1}) \end{aligned}$$

$\sigma_i \sigma_{i+1} \ldots \sigma_n$ については，それぞれのスタックの先頭の状態だけを取り出した $s_{1,i}$, $s_{1,i+1}$, …, $s_{1,n}$ で代表させる．また，$\sigma_{i-1} = s_{1,i-1}\ s_{2,i-1} \ldots s_{i-1,i-1}$ であるとすると，スタックの先頭から k 個を $\sigma_{i-1,k} = s_{1,i-1} s_{2,i-1} \ldots s_{k,i-1}$ と表す．以下に，条件付確率値の近似を示す．

$$\begin{aligned} &P(l_i, a_i, \sigma_i \sigma_{i+1} \ldots \sigma_n \mid \sigma_{i-1}) \\ &\quad \approx P(l_i, a_i, \sigma_i \sigma_{i+1} \ldots \sigma_{i+j-1} \mid \sigma_{i-1,k}) \\ &\quad \approx \begin{cases} P(l_i, a_i, s_{1,i} s_{1,i+1} \ldots s_{1,i+j-1} \mid \\ \quad s_{1,i-1} s_{2,i-1} \ldots s_{k,i-1}), & \text{if } a_i = \textbf{shift} \\ 1, & \text{if } a_i = \textbf{reduce} \end{cases} \\ &\quad \approx P(l_i, s_{1,i} s_{1,i+1} \ldots s_{1,i+j-1} \mid \\ &\qquad s_{1,i-1} s_{2,i-1} \ldots s_{k,i-1}) \end{aligned}$$

上記の近似モデルにおける確率一般化 LR 構文解析を PGLR(j,k) 構文解析と呼び，j は先読みをする状態数を，また，k を現時点でのスタックの先頭からの状態の個数とする．もし，j と k が十分大きければ $P(l_i, a_i, \sigma_i \sigma_{i+1} \ldots \sigma_n \mid \sigma_{i-1})$ と $P(l_i, a_i, s_{1,i} s_{1,i+1} \ldots s_{1,i+j-1} \mid s_{1,i-1} s_{2,i-1} \ldots s_{k,i-1})$ とが等しくなる．

この PGLR(j,k) 構文解析は，厳密な構文解析木を学習するためにこれから到来する先読みとして状態の個数 j とスタックの先頭からの状態の個数 k が必要となることを示している．この j と k は学習する構文解析木と文法から求めることができ，j,k が十分大きくない場合は，構文解析木の生起確率の厳密な値を求めることができない．

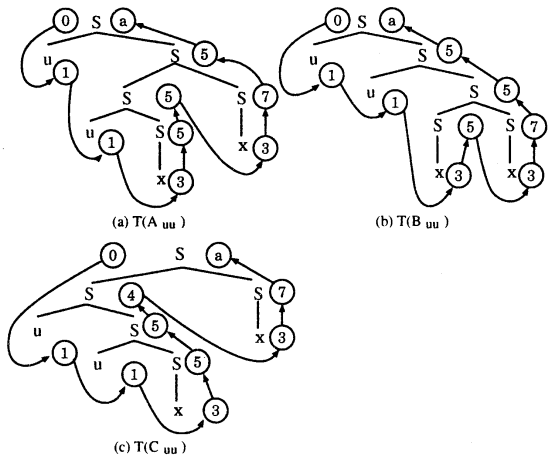

図 1: 入力 $uuxx$ の構文解析木

4.1 PGLR(j,k) 構文解析の例

本節では，PGLR(j,k) 構文解析の例を示す．その例として，次の文法，入力データ，構文解析木の出現個数が与えられるものとする．

(1) **文法**: 文法 $G_1 = (P_1, N_1, T_1, S, \$)$, $P_1 = \{1 : S \to uS, 2 : S \to vS, 3 : S \to x, 4 : S \to SS\}$, $N_1 = \{S\}$, $T_1 = \{u, v, x\}$ を用いて示す．ここで，文法 G_1 に対する LR 表を表 1に示す．

(2) **入力データ**: $uuxx$, $uvxx$, $vuxx$, $vvxx$.

(3) **構文解析木の出現個数**: 入力データから取り出される構文解析木の出現個数を定義する．入力データから作成される構文解析木は $uuxx$ に対して $T(A_{uu})$, $T(B_{uu})$, $T(C_{uu})$, $uvxx$ に対して $T(A_{uv})$, $T(B_{uv})$, $T(C_{uv})$, $vuxx$ に対して $T(A_{vu})$, $T(B_{vu})$, $T(C_{vu})$, $vvxx$ に対して $T(A_{vv})$, $T(B_{vv})$, $T(C_{vv})$ の 12 種類の構文解析木が作成される．この内 $uuxx$ から作成される構文解析木 $T(A_{uu})$, $T(B_{uu})$, $T(C_{uu})$ を図 1 に示す．出現個数は 12 種類の構文解析木 $T(A_{uu})$, $T(B_{uu})$, ..., $T(C_{vv})$ に対応させて，$N_{A_{uu}}$, $N_{B_{uu}}$, ..., $N_{C_{vv}}$ 個出現するものとする．また，出現個数の全数を N, 入力データごとの出現個数を N_{uu}, N_{uv}, N_{vu}, N_{vv} で表し，次のように定義する．

$$N = N_{A_{uu}} + N_{B_{uu}} + N_{C_{uu}} + N_{A_{uv}} + N_{B_{uv}} + N_{C_{uv}} + N_{A_{vu}} + N_{B_{vu}} + N_{C_{vu}} + N_{A_{vv}} + N_{B_{vv}} + N_{C_{vv}},$$
$$N_{uu} = N_{A_{uu}} + N_{B_{uu}} + N_{C_{uu}},$$
$$N_{uv} = N_{A_{uv}} + N_{B_{uv}} + N_{C_{uv}},$$
$$N_{vu} = N_{A_{vu}} + N_{B_{vu}} + N_{C_{vu}},$$
$$N_{vv} = N_{A_{vv}} + N_{B_{vv}} + N_{C_{vv}}.$$

図 1 中の○で囲まれた数字は LR 状態遷移図の状態を表し，状態の遷移の様子を付加している．構文解析木ごとの付加されている状態遷移を次に示す．

A_{uu}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

B_{uu}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

C_{uu}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_4 \overset{\$,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

A_{uv}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

B_{uv}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

C_{uv}: $S_0 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_4 \overset{\$,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

A_{vu}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

B_{vu}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_5 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

C_{vu}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{u,\mathbf{sh}}{\Longrightarrow} S_1 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_5 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_4 \overset{\$,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

A_{vv}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

B_{vv}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_6 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

C_{vv}: $S_0 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{v,\mathbf{sh}}{\Longrightarrow} S_2 \overset{x,\mathbf{sh}}{\Longrightarrow} S_3 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_6 \overset{x,\mathbf{re}}{\Longrightarrow} S_4 \overset{\$,\mathbf{sh}}{\Longrightarrow} S_3 \overset{\$,\mathbf{re}}{\Longrightarrow} S_7 \overset{\$,\mathbf{re}}{\Longrightarrow} S_{acc}$

出現確率のパラメータの先読み状態は，終端記号によって遷移し次の終端記号までに遷移する状態と考えて良い．構文解析木 $T(A_{uu})$ からは，最初の文字 u に対する出現確率 $P(u, S_1 \mid S_0)$, 2 番目の文字 u に対する出現確率 $P(u, S_1 \mid S_0S_1)$, 3 番目の文字 x に対する出現確率 $P(x, S_3S_5S_5 \mid S_0S_1S_1)$, 4 番目の文字 x に対する出現確率 $P(x, S_3S_7S_5S_{acc} \mid S_0S_1S_5)$ が得られる．

表 2は，入力データ全てに対するスタックの状態と先読み状態の出現確率である．表 2から分かるとおり，学習したデータから構文解析木を 1 個に決めるには，時点 i におけるスタックの先頭から 4 個の状態と先読みの 5 個の状態が必要である．よって，この学習例は PGLR(4, 5) 構文解析といえる．

構文解析木 $T(A_{uu})$ の生起確率 $P(T(A_{uu}))$ は，次のように計算される．

表 2: 状態の出現確率

$P(u, S_1 \mid S_0) = (N_{uu} + N_{uv})/N$
$P(v, S_2 \mid S_0) = (N_{vu} + N_{vv})/N$
$P(u, S_1 \mid S_0S_1) = N_{uu}/(N_{uu} + N_{uv})$
$P(v, S_2 \mid S_0S_1) = N_{uv}/(N_{uu} + N_{uv})$
$P(x, S_3S_5S_5 \mid S_0S_1S_1) = N_{A_{uu}}/N_{uu}$
$P(x, S_3S_5 \mid S_0S_1S_1) = N_{B_{uu}}/N_{uu}$
$P(x, S_3S_5S_5S_4 \mid S_0S_1S_1) = N_{C_{uu}}/N_{uu}$
$P(x, S_3S_7S_5S_{acc} \mid S_0S_1S_5) = 1$
$P(x, S_3S_7S_5S_5S_{acc} \mid S_0S_1S_1S_5) = N_{B_{uu}}/N_{B_{uu}} = 1$
$P(x, S_3S_7S_{acc} \mid S_0S_4) = 1$
$P(x, S_3S_6S_5 \mid S_0S_1S_2) = N_{A_{uv}}/N_{uv}$
$P(x, S_3S_6 \mid S_0S_1S_2) = N_{B_{uv}}/N_{uv}$
$P(x, S_3S_6S_5S_4 \mid S_0S_1S_2) = N_{C_{uv}}/N_{uv}$
$P(x, S_3S_7S_6S_5S_{acc} \mid S_0S_1S_2S_6) = N_{B_{uv}}/N_{B_{uv}} = 1$
$P(u, S_1 \mid S_0S_2) = N_{vu}/(N_{vu} + N_{vv})$
$P(v, S_2 \mid S_0S_2) = N_{vv}/(N_{vu} + N_{vv})$
$P(x, S_3S_5S_6S_3 \mid S_0S_2S_1) = N_{A_{vu}}/N_{vu}$
$P(x, S_3S_5 \mid S_0S_2S_1) = N_{B_{vu}}/N_{vu}$
$P(x, S_3S_5S_6S_4 \mid S_0S_2S_1) = N_{C_{vu}}/N_{vu}$
$P(x, S_3S_7S_6S_{acc} \mid S_0S_2S_6) = 1$
$P(x, S_3S_7S_5S_6S_{acc} \mid S_0S_2S_1S_5) = N_{B_{vu}}/N_{B_{vu}} = 1$
$P(x, S_3S_6S_6 \mid S_0S_2S_2) = N_{A_{vv}}/N_{vv}$
$P(x, S_3S_6S_3 \mid S_0S_2S_2) = N_{B_{vv}}/N_{vv}$
$P(x, S_3S_6S_6S_4S_3 \mid S_0S_2S_2) = N_{C_{vv}}/N_{vv}$
$P(x, S_3S_7S_5S_6S_{acc} \mid S_0S_2S_2S_6) = N_{B_{vv}}/N_{B_{vv}} = 1$

表 3: パラメータを削減した状態の出現確率

$P(u, S_1 \mid S_0) = (N_{uu} + N_{uv})/N$
$P(v, S_2 \mid S_0) = (N_{vu} + N_{vv})/N$
$P(u, S_1 \mid S_0S_1) = N_{uu}/(N_{uu} + N_{uv})$
$P(v, S_2 \mid S_0S_1) = N_{uv}/(N_{uu} + N_{uv})$
$P(x, S_3S_5S_5 \mid S_1S_1) = N_{A_{uu}}/N_{uu}$
$P(x, S_3S_5 \mid S_1S_1) = N_{B_{uu}}/N_{uu}$
$P(x, S_3S_5S_5S_4 \mid S_1S_1) = N_{C_{uu}}/N_{uu}$
$P(x, \phi \mid S_1S_5) = (N_{A_{uu}} + N_{A_{uv}})/(N_{A_{uu}} + N_{A_{uv}}) = 1$
$P(x, \phi \mid S_1S_1S_5) = N_{B_{uu}}/N_{B_{uu}} = 1$
$P(x, \phi \mid S_4) = 1$
$P(x, S_3S_6S_5 \mid S_1S_2) = N_{A_{uv}}/N_{uv}$
$P(x, S_3S_6 \mid S_1S_2) = N_{B_{uv}}/N_{uv}$
$P(x, S_3S_6S_5S_4 \mid S_1S_2) = N_{C_{uv}}/N_{uv}$
$P(x, \phi \mid S_1S_2S_6) = N_{B_{uv}}/N_{B_{uv}} = 1$
$P(u, S_1 \mid S_0S_2) = N_{vu}/(N_{vu} + N_{vv})$
$P(v, S_2 \mid S_0S_2) = (N_{A_{vv}} + N_{B_{vv}} + N_{C_{vv}})/(N_{vu} + N_{vv})$
$P(x, S_3S_5S_6S_3 \mid S_2S_1) = N_{A_{vu}}/N_{vu}$
$P(x, S_3S_5 \mid S_2S_1) = N_{B_{vu}}/N_{vu}$
$P(x, S_3S_5S_6S_4 \mid S_2S_1) = N_{C_{vu}}/N_{vu}$
$P(x, \phi \mid S_0S_2S_6) = (N_{A_{vu}} + N_{A_{vv}})/(N_{A_{vu}} + N_{A_{vv}}) = 1$
$P(x, \phi \mid S_2S_1S_5) = N_{B_{vu}}/N_{B_{vu}} = 1$
$P(x, S_3S_6S_6 \mid S_2S_2) = N_{A_{vv}}/N_{vv}$
$P(x, S_3S_6S_3 \mid S_2S_2) = N_{B_{vv}}/N_{vv}$
$P(x, S_3S_6S_6S_4S_3 \mid S_2S_2) = N_{C_{vv}}/N_{vv}$
$P(x, \phi \mid S_2S_2S_6) = N_{B_{vv}}/N_{B_{vv}} = 1$

$$
\begin{aligned}
P(T(A_{uu})) &= P(u, S_1 \mid S_0) \times P(u, S_1 \mid S_0S_1) \\
&\quad \times P(x, S_3S_5S_5 \mid S_0S_1S_1) \\
&\quad \times P(x, S_3S_7S_5S_{acc} \mid S_0S_1S_5) \\
&= \frac{N_{uu} + N_{uv}}{N} \times \frac{N_{uu}}{N_{uu} + N_{uv}} \times \frac{A_{uu}}{N_{uu}} \times 1 \\
&= \frac{A_{uu}}{N}
\end{aligned}
$$

5 PGLR(j,k) 構文解析の j,k の値の削減

4.1 節では PGLR(4, 5) 構文解析の例を示した. しかしながら, この例に見るように, 条件付確率のパラメータを学習した構文解析木に基づいて厳密な値を計算しているため, パラメータ j,k の値が大きくなる傾向にある. そこで, パラメータ j, k の値の削減法を以下に示す.

5.1 同一パラメータ消去によるパラメータ j,k の削減

本節では, 出現確率を表す条件付確率のパラメータを比較し, 構文解析木の生起確率値が変化しない限り同じパラメータを削減する方法である.

(1) **削減 1**: スタックの内容に同じ状態を含んでいる場合, 条件となっている状態を削減する.

例えば, $P(x, S_3S_5S_5 \mid S_0S_1S_1)$, $P(x, S_3S_5 \mid S_0S_1S_1)$, $P(x, S_3S_5S_5S_4 \mid S_0S_1S_1)$ では, $S_0S_1S_1$ が条件になっているが, 他の条件付確率から S_1S_1 に条件を削減する.

(2) **削減 2**: 同じ先読み状態を削減する.

例えば, $P(x, S_3S_7S_5S_6S_{acc} \mid S_0S_2S_2S_6)$ では, $S_0S_2S_2S_6$ となることがないので, それを削減して $P(x, \phi \mid S_0S_2S_2S_6)$ とする. なお, ϕ は空集合を表す.

これらの削減 1,2 を行なうことによって, パラメータ j,k の値を削減した出現確率を表 3に示す. 表 3から分かるように, 先読み状態個数とスタックの内容を読む個数は $j = 3$, $k = 4$ に削減することができ, よって **4.1 節**の例は PGLR(3, 4) 構文解析であるといえる.

5.2 条件の変更によるパラメータ j,k の削減

4 節で示した PGLR(j,k) 構文解析における構文解析木 T の生起確率 $P(T)$ は, 次のように定義している.

表 4: パラメータを削減した状態の出現確率

$P(u,1 \mid S_0) = (N_{uu} + N_{uv})/N$
$P(v,1, \mid S_0) = (N_{vu} + N_{vv})/N$
$P(u,1 \mid S_0S_1) = N_{uu}/(N_{uu} + N_{uv})$
$P(v,1 \mid S_0S_1) = N_{uv}/(N_{uu} + N_{uv})$
$P(x,3 \mid S_1S_1) = N_{A_{uu}}/N_{uu}$
$P(x,2 \mid S_1S_1) = N_{B_{uu}}/N_{uu}$
$P(x,4 \mid S_1S_1) = N_{C_{uu}}/N_{uu}$
$P(x,0 \mid S_1S_5) = (N_{A_{uu}} + N_{A_{uv}})/(N_{A_{uu}} + N_{A_{uv}}) = 1$
$P(x,0 \mid S_1S_1S_5) = N_{B_{uu}}/N_{B_{uu}} = 1$
$P(x,0 \mid S_4) = 1$
$P(x,3 \mid S_1S_2) = N_{A_{uv}}/N_{uv}$
$P(x,2 \mid S_1S_2) = N_{B_{uv}}/N_{uv}$
$P(x,4, \mid S_1S_2) = N_{C_{uv}}/N_{uv}$
$P(x,0 \mid S_1S_2S_6) = N_{B_{uv}}/N_{B_{uv}} = 1$
$P(u,1 \mid S_0S_2) = N_{vu}/(N_{vu} + N_{vv})$
$P(v,1 \mid S_0S_2) = (N_{A_{vv}} + N_{B_{vv}} + N_{C_{vv}})/(N_{vu} + N_{vv})$
$P(x,4 \mid S_2S_1) = N_{A_{vu}}/N_{vu}$
$P(x,2 \mid S_2S_1) = N_{B_{vu}}/N_{vu}$
$P(x,4 \mid S_2S_1) = N_{C_{vu}}/N_{vu}$
$P(x,0 \mid S_0S_2S_6) = (N_{A_{vu}} + N_{A_{vv}})/(N_{A_{vu}} + N_{A_{vv}}) = 1$
$P(x,0 \mid S_2S_1S_5) = N_{B_{vu}}/N_{B_{vu}} = 1$
$P(x,3 \mid S_2S_2) = N_{A_{vv}}/N_{vv}$
$P(x,3 \mid S_2S_2) = N_{B_{vv}}/N_{vv}$
$P(x,5 \mid S_2S_2) = N_{C_{vv}}/N_{vv}$
$P(x,0 \mid S_2S_2S_6) = N_{B_{vv}}/N_{B_{vv}} = 1$

$$P(T) \approx \prod_{i=1}^{n-1} P(l_i, s_{1,i}s_{1,i+1}\cdots s_{1,i+j-1} \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1})$$

上記の生起確率 $P(T)$ から入力 l_i によって求められる構文解析木の部分木の生起確率は，$P(l_i,\ s_{1,i}s_{1,i+1}\ \cdots\ s_{1,i+j-1} \mid s_{1,i-1}\ s_{2,i-1}\ \cdots\ s_{k,i-1})$ で求められる．先読み状態のそれぞれの状態には $s_{1,i}, s_{1,i+1}, \ldots, s_{1,i+j-2} \in \mathbf{S_r}$, $s_{1,i+j-1} \in \mathbf{S_s}$ が成り立つ．例えば，表 3中の出現確率 $P(x, S_3S_5S_5S_4 \mid S_1S_1)$ の先読み状態のそれぞれの状態は $S_3, S_5, S_5 \in \mathbf{S_r}$, $S_4 \in \mathbf{S_s}$ である．また，このことは，図 1の $T(C_{uu})$ の 3 文字目の入力である x から次の x までの状態遷移リストからも分かる．

そこで，動作を条件に加えるかわりに，各状態の先読み状態中における位置を条件とする出現確率の積によって構文解析木の部分木の生起確率を表すことにする．

$$P(l_i, s_{1,i}s_{1,i+1}\cdots s_{1,i+j-1} \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1}) = (\prod_{l=0}^{j-2} P(l_i \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1}, a_{i+l}))\times P(l_i, j, a_{i+j-1} \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1})$$

$a_i, a_{i+1}, \ldots, a_{i+j-2}$ = **reduce**, a_{i+j-1} = **shift**.

また，動作 a_{i+l} = **reduce**, $l = 0, \ldots, j-2$ では，$P(\ l_i \mid s_{1,i-1}\ s_{2,i-1}\ \cdots\ s_{k,i-1},\ a_{i+l}\ ) = 1$ であるので，構文解析木 T の生起確率 $P(T)$ は次のように定義できる．

$$P(T) \approx \prod_{i=1}^{n-1} P(l_i, j, \mathbf{shift} \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1}) \approx \prod_{i=1}^{n-1} P(l_i, j \mid s_{1,i-1}s_{2,i-1}\cdots s_{k,i-1})$$

表 4に先読み状態の位置による出現確率を示す．

4.1 節の例における構文解析木 $T(C_{uu})$ に対する生起確率 $P(T(C_{uu}))$ は，次のように計算される．

$$\begin{aligned} P(T(C_{uu})) &= P(u,1 \mid S_0) \times P(u,1 \mid S_0S_1) \times P(x,4 \mid S_1S_1) \times P(x,0 \mid S_4) \\ &= \frac{N_{uu}+N_{uv}}{N} \times \frac{N_{uu}}{N_{uu}+N_{uv}} \times \frac{C_{uu}}{N_{uu}} \times 1 \\ &= \frac{C_{uu}}{N} \end{aligned}$$

6 むすび

本稿では，条件付確率のパラメータを見直すことによって，厳密な確率一般化 LR 構文解析ができることを示した．ただし，パラメータの数が多くなるようなモデルでは，読み込む状態数を減らすことによって近似解を求めることも考えられる．

参考文献

[1] M. Tomita, *An Efficient Augmented - Context-Free Parsing Algorithm*, Computational Linguistics, 13,1-2, pp31-46, 1987.

[2] T. Briscoe and J. Carroll, *Generalized Probabilistc LR Parsing of Natural Language(Corpora) with Unification-Based Grammars*, Computational Linguistics, Vol.19, No.1, pp.25-59, 1993.

[3] E. Charniak, *Statistical langugae learning*, MIT press, 1993.

[4] 井上 謙蔵, コンパイラ, 丸善, 1994.

[5] K. Inui, V. Sornlertlamvanich, H. Tanaka and T. Tokunaga, *Probabilistic GLR Parsing: A New Formalization and Its Impact of Parsing Performance*, Journal of Natural Language Processing, Vol. 5, No. 3, pp.33-52, 1998.